Journal of Research in Science Teaching

The Relative Effects and Equity of Inquiry-Based and Commonplace Science Teaching on Students' Knowledge, Reasoning and Argumentation

Journal:	*Journal Research Science Teaching*
Manuscript ID:	JRST-2008-11-0333.R3
Wiley - Manuscript type:	Research Article
CONTENT AREA:	biology
FOCUS AREA:	achievement, inquiry, equity
GRADE LEVEL:	secondary
METHODOLOGY/APPROACHES:	experimental, classroom research

Title:

The Relative Effects and Equity of Inquiry-Based and Commonplace Science Teaching on Students' Knowledge, Reasoning and Argumentation

Running Head:

EFFECTS OF INQUIRY-BASED AND COMMONPLACE TEACHING

Authors:

Christopher D. Wilson, Joseph A. Taylor, Susan M. Kowalski, and Janet Carlson

Contact Details:

Christopher D. Wilson
BSCS Center for Research & Evaluation
5415 Mark Dabling Blvd.
Colorado Springs
CO 80918
cwilson@bscs.org
719 219 4106
FAX: 719.531.9104

Susan M. Kowalski
BSCS Center for Research & Evaluation
5415 Mark Dabling Blvd.
Colorado Springs
CO 80918
skowalski@bscs.org
719 531 5550 ex 148
FAX: 719.531.9104

Joseph A. Taylor
Director,
BSCS Center for Research & Evaluation
5415 Mark Dabling Blvd.
Colorado Springs
CO 80918
jtaylor@bscs.org
719 219 4104
FAX: 719.531.9104

Janet Carlson
Executive Director, BSCS
5415 Mark Dabling Blvd.
Colorado Springs
CO 80918
jcarlson@bscs.org
719 531 5550 ex 126
FAX: 719.531.9104

Abstract

We conducted a laboratory-based randomized control study to examine the effectiveness of inquiry-based instruction. We also disaggregated the data by student demographic variables to examine if inquiry can provide equitable opportunities to learn. Fifty-eight students aged 14–16 years old were randomly assigned to one of two groups. Both groups of students were taught toward the same learning goals by the same teacher, with one group being taught from inquiry-based materials organized around the BSCS 5E instructional model, and the other from materials organized around commonplace teaching strategies as defined by national teacher survey data. Students in the inquiry-based group reached significantly higher levels of achievement than students experiencing commonplace instruction. This effect was consistent across a range of learning goals (knowledge, reasoning, and argumentation) and time frames (immediately following the instruction and four weeks later). The commonplace science instruction resulted in a detectable achievement gap by race, whereas the inquiry-based materials instruction did not. We discuss the implications of these findings for the body of evidence on the effectiveness of teaching science as inquiry; the role of instructional models and curriculum materials in science teaching; addressing achievement gaps; and the competing demands of reform and accountability.

Keywords: inquiry, equity, achievement, biology.

The Relative Effects and Equity of Inquiry-Based and Commonplace Science Teaching on Students' Knowledge, Reasoning and Argumentation

From Dewey to the present, inquiry has been an increasingly prominent theme in multiple science education reform movements worldwide. However, the transition from theory and advocacy to practice and policy has been unsatisfactory. The paradox of educational reform without change is not exclusive to the sciences (Cuban, 1988; Woodbury & Gess-Newsome, 2002), but it is nevertheless surprising that such a sustained and largely consistent drive for reform has had such little impact on teacher practice. Two large scale studies from Horizon Research, Inc. (Weiss, Pasley, Smith, Banilower, & Heck, 2003; Hudson, McMahon, & Overstreet, 2002) highlight the uncommonness of inquiry-based teaching in the United States. From classroom observations and interviews with 364 science and mathematics teachers, Weiss et al. (2003) found that inquiry was a focus of only 2 percent of science lessons in grades 9–12. This finding mirrors those in a survey of 5,278 teachers (Hudson et al., 2002) in which teaching practices and student objectives characteristic of inquiry consistently occurred with less frequency and emphasis than traditional teaching methods and learning goals. Inquiry is a central theme in the National Science Education Standards (NSES; National Research Council [NRC], 1996) and its clarifying documents (NRC, 2000) as well as in significant international reform documents (Osborne & Dillon, 2008; Australian Education Council, 1994; Tomorrow 98, 1992; Ministry of Education, 1999). In the U.S. only 12 percent of high school science teachers in the Hudson et al. survey said that they had "implemented recommendations from the National Education Standards in [their] science teaching" to a great extent and only 4 percent strongly agreed with the statement "I am prepared to explain the NRC National Science Education

Standards to my colleagues." The infrequency of inquiry-based teaching found in these large-scale surveys and interviews is consistent with the findings of studies from the full range of research traditions (R. Anderson, 2002; Abd-El-Khalick et al., 2004; Crawford, 2007), as well as data collected in countries other than the U.S. (Osborne, 2009).

Many barriers to implementing inquiry in a manner consistent with the vision of the NSES have been described in the literature (Welch, Klopfer, & Aikenhead, 1981; Gallagher, 1989; Roehrig & Luft, 2004; Lederman, 2004; McGinnis, Parker, & Graeber, 2004; and Crawford, 2007). R. Anderson (2002) categorizes these as *political dilemmas* (such as parental resistance and conflicts between teachers), *cultural dilemmas* (such as differing beliefs and values about learning and assessment), and *technical dilemmas* (which include limited abilities to teach and assess). Similarly, Tobin and McRobbie (1996) describe a series of *cultural myths* - beliefs about teaching and learning that constrain teachers' pedagogical moves and result in teaching practices discordant with teaching science as inquiry (Lotter, Harwood and Bonner, 2007). In recognizing these dilemmas and myths, meeting the demands of an age of reform presents a significant challenge, but we are also in an age of accountability that has brought its own obstacles to teaching science as inquiry. The No Child Left Behind [NCLB] legislation (U.S. Department of Education, 2002) and the associated accountability movement have led to an increased emphasis on standardized testing to measure teacher and school effectiveness. In turn, some have argued (see for example Blanchard, Annetta, & Southerland, 2008) that standardized testing (a) has resulted in teaching practices that are at odds with those advocated in the national science education reform documents (American Association for the Advancement of Science (AAAS), 1993, 2000; NRC, 1996, 2000), (b) has had negative effects on science teachers' perceptions of the quality of their teaching (Shaver, Cuevas, Lee, & Avalos, 2006;

Southerland, Abrams, & Hutner, 2007), and (c) has created pressures for teachers to prepare students for tests that cover large amounts of content and emphasize factual knowledge (Whitford & Jones, 2000). NCLB and the current climate in the U.S. therefore present one further obstacle to inquiry's role in reform: accountability and inquiry-based teaching can appear incompatible to teachers (Blanchard et al., 2008). We explore this (perhaps false) dichotomy in this study by examining the achievement of students who receive instruction guided by inquiry-based curriculum materials, and students who receive instruction toward the same learning goals guided by materials designed around commonplace teaching practices.

While NCLB and the associated accountability movement have changed how states assess teacher and school effectiveness, they have also resulted in a shift in the expectations for evidence in education research. Federal policies have begun to advocate *evidence-based reform*—in which the adoption of programs or practices is based on rigorous research conducted with methods derived from the medical and natural sciences, particularly experiments in which subjects are randomly assigned to treatments (Slavin, 2008). To ensure that there was no doubt about its significance, the use of "scientifically-based research" to inform policy decisions regarding education programs and practices was mentioned in the No Child Left Behind Act (U.S. Department of Education, 2002) more than 100 times (Slavin, 2008). The U.S. Department of Education has also championed efforts to synthesize research findings related to effective programs and practices, including the What Works Clearinghouse (WWC), the Best Evidence Encyclopedia (BEE), and the (now defunct) Comprehensive School Reform Quality Center (CSRQ). Each of these centers assesses the quality of research studies primarily by their methodological rigor, with the highest ratings going to studies incorporating randomized experiments. We are therefore met with a challenge. If, within the current climate of accountability and evidence-based reform in the U.S., the cumulative vision of a century of science education reform is to become commonplace practice, the question becomes: *What is the evidence that demonstrates the effectiveness of inquiry-based materials and teaching?*

The Evidence on the Effectiveness of Inquiry-Based Materials and Teaching

The science education community has published a wide range of findings about inquiry-based teaching and learning including inconclusive, mixed, or negative results (see Colburn, 2008 for a review). The most significant challenges have come recently from cognitive scientists. One prominent example is the provocatively titled article *Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching*, in which Kirshner, Sweller, and Clark (2006) review a small number of studies that, they argue, provide evidence against the effectiveness of inquiry-based materials and teaching. The studies they reviewed include some that showed how pure discovery teaching methods can lead to frustration (Hardiman, Pollatsek, & Weil, 1986; Brown & Campione, 1994), some that showed how discovery learning is inefficient because it can lead to false starts (Carlson, Lundy, & Schneider, 1992; Schauble, 1990), and some that found support for direct instruction over discovery learning (Moreno, 2004). The title of their study suggests that Kirshner et al. (2006) equate inquiry with other instructional approaches as being characterized by "minimal guidance during instruction," an assertion contested in a response by Hmelo-Silver, Duncan, and Chinn (2007).

Hmelo-Silver et al. (2007) describe research on the many forms of scaffolding involved in inquiry-based teaching (Collins, Brown, & Newman, 1989; Golan, Kyza, Reiser, & Edelson, 2002; Jackson, Stratford, Krajcik, & Soloway, 1996) and firmly disassociate it from the discovery learning examined in the studies cited by Kirshner et al. (2006). Hmelo-Silver et al. (2007) describe how inquiry is not only far from being "minimally guided," but in fact relies on significant scaffolding to guide student learning, and commonly involves timely direct instruction (Bybee et al., 2006; Krajcik, Czerniak, & Berger, 1999; Schmidt, 1983; Schwartz & Bransford, 1998).

Consistent with their lack of distinction between different instructional philosophies/models, Kirshner et al. (2006) highlight the work of Klahr and Nigam (2004) as providing particularly significant evidence against inquiry-based materials and teaching, because the authors "not only tested whether science learners learned more via a discovery versus direct instruction route but also, once learning had occurred, whether the quality of learning differed." The work by Klahr and colleagues (Chen & Klahr, 1999; Klahr & Nigam, 2004) has indeed stimulated review and discussion of the relative importance of direct instruction and discovery learning as instructional approaches for science teaching, but in neither article do the authors make any claims about inquiry. Furthermore, the authors' operational definition of *direct instruction* in these studies has been shown by Bybee et al. (2006) to incorporate many aspects of an inquiry-based instructional model, and their operational definition of *discovery learning* has been shown by Blanchard et al. (2008) to involve no teacher scaffolding. Consequently, the work of Klahr and colleagues shows little resemblance to how inquiry is described in the NSES (NRC, 1996, 2000) or to guided inquiry (Colburn, 2000) or level 2 inquiry (Schwab, 1962). Finally, in a study examining acquisition of the same learning goal as Klahr & Nigam (2004) by different instructional approaches, Dean and Kuhn (2006) found that direct instruction was "neither a necessary nor sufficient condition for robust acquisition or for maintenance over time." Despite these alternative interpretations of the instructional approaches in the Klahr et al. studies, the implications of their research have been stated in the extreme by the popular press. Unfortunately, characterization of the instructional approaches of discovery and direct instruction as diametrically opposed options, rather than as part of a set of strategies that may be integrated carefully in the same science classroom, has done a disservice to both approaches.

In their response to Kirshner et al. (2006), Hmelo-Silver et al. (2007) concede that experimental studies of inquiry-based materials and teaching are limited, yet they do cite a small number of experimental and quasi-experimental studies that compare inquiry-based teaching to other instructional approaches. These include a study by Hickey, Kindfeld, Horwitz, and Christie (1999) that found that students using the inquiry-based GenScope™ learning environment showed significantly higher learning gains than students in comparison classrooms that did not incorporate inquiry-based strategies and materials. Using performance on state standardized tests as the measure of student learning, Geier et al. (2008) found significantly higher pass rates among urban middle school students using inquiry-based materials compared to students using traditional materials. The effects were both cumulative (more exposure to inquiry-based units resulted in higher achievement on the tests) and enduring (the learning gains were evident a year and half after participation in the units). Hmelo-Silver et al. (2007) also describe a study by Lynch, Kuipers, Pyke, and Szesze (2005) in which students receiving inquiry-based instruction outperformed students in comparison groups, regardless of ethnicity, socioeconomic status, gender, and ESOL status. Hmelo-Silver et al. (2007) conclude: “there is growing evidence from large-scale experimental and quasi-experimental studies demonstrating that inquiry-based instruction results in significant learning gains in comparison to traditional instruction.”

There is a long history of research into inquiry-based teaching and curriculum materials that involves research designs that are not experimental. Two classic meta-analyses looked across studies examining various curriculum materials and teaching strategies, and both found substantial effect sizes for student learning in favor of an inquiry-based approach (Shymansky, Kyle, & Alport, 1983; Wise & Okey, 1983). Colburn (2008) provides a review of studies examining the effectiveness of inquiry-based teaching up to the mid 1990s. Some notable studies discussed by Colburn include work by Westbrook and Rogers (1994), who examined the effectiveness of instruction organized around different learning cycle models. They found significant gains in students' reasoning abilities only after instruction organized around learning cycles that closely resembled guided and open inquiry (see Lawson, 1995, for a comprehensive review of other studies examining inquiry-based learning cycles). Colburn (2008) also describes the work of Leonard (1983) who compared college-level instruction with inquiry-based materials and with instruction from traditional materials. Studies by Leonard (1983), Hall and McCurdy (1990), and Leonard, Cavana, and Lowery (1981) found significant learning gains when students were taught using inquiry-based materials. Similarly to the Hmelo-Silver et al. (2007) conclusion, Colburn (2008) notes: "Most studies I examined supported the collective conclusion that inquiry-based instruction was equal or superior to other instructional models for students producing higher scores on content achievement tests."

Finally, recent studies by Blanchard et al. (2008), Lederman, Lederman, and Wickman (2008) and Lewis and Lewis (2008) shine further light on questions regarding the effectiveness of inquiry-based curriculum materials and teaching strategies. Blanchard et al. (2008) compared learning gains in middle and high school students after being taught a forensic unit by either inquiry-based or traditional approaches. Their study, involving 1,800 students and 24 teachers from seven schools, showed significantly higher posttest scores among the students taught by a guided inquiry approach, as compared to students taught by traditional methods. Lederman et al.

(2008) conducted a study with teachers in Sweden and the United States, in which teachers taught units either by direct instruction, guided inquiry, or a hybrid of the two. While the mixed approach was the most successful with respect to increasing subject matter knowledge, knowledge of scientific inquiry, as well as attitudes towards science, the differences were not statistically significant for any of the approaches. Lewis and Lewis (2008) extend the above K-12 findings to the college level, where undergraduate students taught via peer-led guided inquiry achieved significantly higher academic performance across multiple measures than students taught using a traditional pedagogical approach.

Rationale for the Study

From the perspective of the evidence-based reform movement, the evidence for the effectiveness of inquiry-based materials and teaching to date can only be seen as inconclusive. In this study, we address this ambiguity by employing the methods of scientifically-based research (Slavin, 2008; Shavelson & Towne, 2002). Specifically, we designed a study to examine the differences between the achievement of students who received instruction guided by an inquiry-based unit organized around the BSCS 5E Instructional Model and students who received instruction on the same content based on an instructional unit designed around commonplace teaching practices as defined by national surveys. We are therefore studying the effectiveness of the enactment of inquiry-based and commonplace materials, which is influenced by the teacher, the students, and the curriculum materials themselves (Remillard, 1999, 2005). From this point on when we use the term *instruction*, we are referring to the enacted curriculum.

Because significant achievement gaps by gender, race/ethnicity, and socioeconomic status remain in the U.S. (Clewell & Campbell, 2002) despite the long-standing call for science for all Americans, we disaggregated data by various student demographic variables to examine if inquiry-based instruction can provide equitable opportunities to learn. As described below, we use the Horizon Research, Inc. survey and interview data (Weiss et al., 2003; Hudson et al., 2002) to operationally define *commonplace instruction*, and use the BSCS 5E Instructional Model (Bybee, 1997; Bybee, Carlson Powell, & Trowbridge, 2007; Bybee & Landes, 1990; Bybee et al., 2006) to organize the inquiry-based unit.

Inquiry and the BSCS 5E Instructional Model

Since the late 1980s, BSCS has used one instructional model extensively in the development of new curriculum materials and professional development experiences (Bybee et al., 2006). That model is commonly referred to as the BSCS 5E Instructional Model, or the BSCS 5Es, and consists of the following phases: engage, explore, explain, elaborate, and evaluate. Each *E* supports classroom experiences and teaching strategies that provide students with opportunities to construct content understanding within the context of experiences consistent with science as inquiry. Once internalized, the model also informs the many instantaneous decisions that science teachers must make in classroom situations.

While the BSCS 5Es and inquiry are not synonymous, the former represents an instructional model based on constructivist theories of learning that provides strong guidance and support for an approach to teaching that promotes student inquiry. As shown in Table 1, each of the five essential features of inquiry (NRC, 2000) is represented at various stages of the BSCS 5E Instructional Model. Further, the BSCS 5Es outline student and teacher roles that are consistent with the NSES *Content Standards for Scientific Inquiry* and the inquiry-based *Science Teaching Standards* respectively (NRC, 1996, 2000).

Measuring Learning

We have described above how differences in the way researchers define inquiry can lead to difficulties in comparing instructional approaches. While there is certainly a lack of agreement among researchers and practitioners regarding the meaning of inquiry-based instruction (Minstrell, 2000; Barman, 2002; Lederman, 2003), the multiple understandings and abilities of inquiry described in the NSES (NRC, 1996) and other documents (NRC, 2000) lead to a number of possible learning outcomes. As such, different aspects of student learning measured in studies examining inquiry include students' mastery of subject matter, scientific reasoning, understanding of the nature of science, interest and attitudes toward science, and various science skills. Any study examining the effectiveness of inquiry-based instruction must therefore be careful in ensuring that their measures of effectiveness are clearly aligned with specific learning goals. Additionally, the types of items used to assess the effectiveness of inquiry-based instruction have been as varied as the effects they measure. Blanchard et al. (2008) express concern about how different item formats may favor different types of learning gains in inquiry-based or traditional instruction groups. These concerns are echoed by Shymansky, Yore, Annetta & Everett (2008), who question whether multiple-choice tests allow students to reveal content-related problem-solving or critical-thinking skills, rather than just knowledge of facts and vocabulary.

In this study we measure three goals of science education that are reflected in the foci of prominent national and international science education documents (Rutherford & Ahlgren, 1989; AAAS, 1993; NRC, 1996, 2000; Bransford, Brown, & Cocking, 1999; Osborne & Dillon, 2008; Australian Education Council, 1994; Tomorrow 98, 1992; Ministry of Education, 1999) as well as in many reform-based curriculum materials. We included multiple outcomes and measures to reflect both the multiple learning goals of inquiry-based instruction, as well as to avoid bias caused by the measures being unfairly aligned with the goals and procedures of the treatment group (Briggs, 2008; Schoenfeld, 2006; Confrey, 2007). The three measured outcomes are:

- *Scientific knowledge.* This construct reflects both the foundation of factual knowledge required to develop competence in an area of inquiry (Bransford et al., 1999) as well as the common focus of science instruction on factual recall, scientific vocabulary, and assessments with clearly right and wrong answers (Driver, Newton, & Osborne, 2000). As such, this outcome is measured with dichotomous true/false and multiple-choice items.
- *Scientific reasoning through application of models.* Bransford et al. (1999) describe the need for students to organize scientific ideas in the context of a conceptual framework. Such organizing structures can be seen as analogous with the scientific models described by Gilbert, Boulter, and Rutherford (1998), Geire (1999), and Cartier, Rudolph, & Stewart (2001). One measure of students' understanding of scientific models is their ability to apply them to reasoning about new patterns and data in new contexts (Anderson, 2003). Here we measured students' ability to reason with scientific models through constructed-response items in which students are asked to explain or predict patterns in novel situations. We scored their responses along a continuum representing increasingly sophisticated accounts, ranging from informal cultural models, to scientific models that traverse physiological, organismal, and environmental scales.
- *Construction and critique of scientific explanations.* The NRC *Standards* (NRC, 1996) and AAAS *Benchmarks* (AAAS, 1993) both emphasize developing and evaluating

scientific explanations (often referred to as argumentation)—practices argued to be more representative of the social practice of science than those found in traditional science teaching and learning (Driver et al., 2000). In this study, students' ability to construct and critique arguments was assessed via standardized open-ended interviews, in which students were asked to develop explanations for patterns in given data, as well as critique given explanations for those patterns. The interviews were scored according to a modified version of the McNeill, Lizotte, Krajcik, and Marx (2006) Claim-Evidence-Reasoning framework (which in turn is an adaptation of Toulmin's argumentation model; Toulmin, 1958).

With respect to these outcomes, our primary research question was: What is the effectiveness of inquiry-based materials on student learning as compared to commonplace materials? With this question being broken into the following sub questions:

a) To what extent can differences in student achievement between the inquiry-based and commonplace groups be attributed to randomized group assignment?
b) Does student race/ethnicity, gender, or socioeconomic status account for variation in posttest scores above and beyond variation accounted for by pretest scores and group assignment?
c) What differences in achievement by treatment group exist specific to the learning goals of knowledge, reasoning, and argumentation?

Methods

Experimental Design and Student Sample

Since one goal of this study was to investigate whether causal inferences could be made about the effectiveness of inquiry-based curriculum instruction, a laboratory-based randomized control design was used. An invitation was sent to Colorado Springs area schools, youth organizations, and home-school groups inviting children aged between 14 and 16 years to participate in a research study involving 14 hours of instruction and testing over the course of two weeks in the summer. Sixty students were successfully recruited, and each was randomly assigned via a coin flip to either a group that would receive inquiry-based instruction based on curriculum materials organized around the BSCS 5Es or a group that would receive instruction on the same content but organized around commonplace teaching practices.

The 58 study participants came from 24 schools from seven districts from across a range of urban, suburban, and rural areas; five of the students attended private schools and two were home-schooled. With respect to gender, race, age, and free/reduced lunch status, no significant differences were found in the composition of each of the two treatment groups. Table 2 summarizes these data. Each student received compensation at the end of the data collection as long as she or he attended all class sessions, completed all pretests and posttests, and participated in a standardized open-ended interview four weeks after the classes. The students were unaware of the purpose of the study, their group assignment, and as much as was possible, the existence of the other treatment group. To remove the possibly confounding effects of multiple teachers, both units were taught by the same teacher in a controlled laboratory setting in the BSCS classroom in Colorado Springs. The teacher selected for this study had 27 years of experience teaching in public schools, a Ph.D. in curriculum and instruction, and experience teaching with a wide range of traditional and inquiry-based materials.

Unit Development

The instructional unit selected for this study was *Sleep, Sleep Disorders, and Biological Rhythms* from the National Institutes of Health (NIH) Curriculum Supplement Series (BSCS, 2003). This unit was selected because (a) the content covered in the unit falls outside of the regular K-12 curriculum, and so would be largely unfamiliar to all students; (b) the length of unit fit within the study's constraints; and (c) the unit was already designed to be inquiry-based within the framework of the BSCS 5Es. The original sleep unit contained a pre-unit sleep diary and five lessons covering topics including circadian rhythms and the biological clock, physiological changes during sleep, and the science of sleep disorders. The original NIH sleep unit was modified for the purposes of this study to produce two new instructional units that exemplified commonplace teaching and inquiry-based teaching as described below.

The Commonplace Unit. The two research documents from Horizon Research, Inc. described previously (Weiss et al., 2003; Hudson et al., 2002) were used to help establish commonplace teaching practice. Items from the Hudson et al. (2002) survey that were particularly useful for defining *commonplace* included those that examined:

1. The emphasis given to various instructional objectives, such as learning terms and facts, learning to evaluate scientific arguments, or learning about the nature of science.
2. The frequency of teachers' use of various instructional strategies, such as introducing content through formal presentations, posing open-ended questions, or asking students to consider alternative explanations.

3. The frequency of student participation in various activities, such as watching a demonstration, following specific instructions in an activity or investigation, or designing and implementing their own investigation.

Some useful items from the Weiss et al. (2003) classroom observations and interviews included those that examined:

1. The percent of time students spend as a whole class, in small groups, and as individuals.
2. The frequency of activities in science lessons, such as teacher lectures, students doing hands-on activities, and students completing textbook or worksheet problems.
3. The content focus of the observed lessons, including if science as inquiry was a focus of the lessons.

Each of the lessons in the original NIH sleep unit was modified to reflect the frequency of teaching practices illustrated by patterns in the data from the Horizon Research, Inc. studies. To reflect commonplace practices, changes were also made to the order of the lessons, as well as to the connections between the lessons. Rather than merely focusing on didactic approaches to teaching, the commonplace unit included strategies and activities such as group work and experiments in the same frequency as the survey and interview data.

The Inquiry-Based Unit. Despite the original NIH sleep unit being organized around the BSCS 5Es, the unit was reviewed to insure consistency with teaching science as inquiry within the BSCS 5E model. A small number of changes were made to more fully represent the BSCS

5E Instructional Model and the processes of scientific inquiry. These changes included the following:

- Adding some explicit scaffolding to one of the inquiry activities.
- Moving an activity from lesson 3 to lesson 1 to serve as an *Engage* activity.
- Focusing the *Explore* activity on students finding patterns and negotiating those with their peers, not drawing conclusions.
- Emphasizing the negotiation of explanations between students, alternative hypotheses, and evidence-based arguments.
- Writing more discussion questions, including probes, for the teacher to use that extended the opportunity for students to develop explanations.

Both sets of materials were reviewed and revised by expert curriculum developers to insure that while the instructional approaches differed, the learning goals remained the same. Table 3 summarizes differences between the key student activities in each of the five lessons.

Data Collection

Pretest, Posttest and Interview. All students completed a pretest and posttest immediately before and after instruction and participated in a thirty-minute interview four weeks following the unit. The pretest and the posttest were identical, and contained four multiple-choice items, eight true/false items, and five constructed response items. The true/false and multiple choice items were designed to focus on simple "facts" and vocabulary contained within the sleep unit, while the constructed response items required students to apply scientific models of sleep behavior to reasoning about data presented in new contexts. The final test items were selected from a larger pool of items by content experts, and underwent field testing with students not participating in the study and subsequent refinement. Students completed the pencil and paper tests in controlled conditions. The maximum score on both the pretest and posttest was 74 points, with 24 points coming from the 12 true/false and multiple choice items, scored 2 points each, and 50 points from the five constructed response items. The mean item difficulty for the true/false and multiple choice items was 0.789, with the total test having a reliability index of 0.695 (Cronbach's alpha). All items had a positive discrimination index.

A thirty-minute standardized open-ended interview protocol was developed around the topics of sleep behavior, circadian rhythms, and the biological clock. During these interviews, students were presented with sleep data in the form of actograms—representations of sleep behavior that were not used during either instructional unit. Based on the data in the actograms, students were guided through the construction of explanations that included environmental and physiological explanations for the observed data, asked for alternative explanations for their observations, and asked to critique given explanations for the patterns in the data. Each interview was recorded on video. The Appendix contains example questions from the pretest/posttest and standardized interview.

Measures of Differences between the Enactments of the Two Units. Each class session was observed by three external researchers, who took comprehensive notes and completed the Reformed Teaching Observation Protocol (RTOP; Piburn et al., 2000; Sawada et al. 2002) for each unit. The teacher also took extensive notes after each lesson, recording his pedagogical moves and differences between his teaching in the two units. Each class session was recorded on video. At the end of the unit, all students completed a survey containing a subset of 17 items from the Constructivist Learning Environment Survey (CLES; Taylor & Fraser, 1991).

The RTOP scores for the inquiry-based unit were significantly higher than those for the commonplace unit across [$t(48) = 9.937$, $p < 0.01$] and within RTOP subscales ($p < 0.01$ for each). Similarly, the mean CLES scores for the inquiry-based unit were significantly higher than the mean scores for the commonplace unit [$t(55) = 3.195$, $p < 0.01$]. Since both high RTOP and CLES scores reflect a classroom environment in which the inquiry-based teaching standards in the NSES (NRC, 1996, 2000) are put in practice, these findings demonstrate that the enactment of the two units reflected the design of the curriculum materials in making a distinction between commonplace and inquiry-based teaching and learning.

Using video recordings of the classroom sessions, we also coded the classes using the 5-minute observation section of the Collaboratives for Excellence in Teacher Preparation (CETP) Core Observation Protocol (Lawrenz , Huffman & Gravely, 2007). These data are presented in Table 4. Researchers were blinded to treatment group (inquiry or commonplace) and assigned codes for each five minute segment of each class for classroom activity, student engagement level, and the level of cognitive demand placed on students. The researchers jointly scored a selection of videos until agreement on coding was reached. Differences were resolved through discussion. When both researchers were comfortable with the process, the remaining videos were scored individually. Multiple codes could be assigned for each five minute segment for classroom activity, provided that the activities occurred simultaneously. For example, it was common to code a segment as both teacher interacting with students and small group discussion; whereas lecture and small group discussion did not tend to be occur in the same five minutes.

Engagement level indicates the percentage of students that were “on task,” doing what they were supposed to be doing. It does not purport to measure student excitement or enthusiasm.

Key differences between classes appear in several classroom activity codes: time spent on lecture was higher for the commonplace group; time spent demanding a higher level of cognitive activity from students was higher in the inquiry group. Furthermore, the Inquiry group spent more time in small group discussions, writing work, and experienced greater teacher-student interactions. In these analyses, note-taking was not considered written work. Rather, written work involved students answering questions, designing experiments, analyzing data, or solving problems in their notebooks. It should be noted that much of the writing work in the inquiry group was connected with *constructing understanding* (e.g. developing an explanation), whereas the students in the commonplace group were usually *receiving knowledge* (completing a worksheet) while writing. Time spent on concepts mapped fairly closely for the two classes with some exceptions. The inquiry group spent some time on how to write scientific questions and how to write a scientific procedure, whereas the commonplace group did not. The commonplace group, instead, spent more time on sleep cycles, measuring sleep cycles, and the astronaut problem (a problem requiring students to analyze data from astronauts to determine if the astronauts were asleep or awake; and if asleep, which stage of sleep). The fact that no time was spent on hands-on activity/materials might appear strange given the common appearance of inquiry-based lessons. However, because the focus of the unit was biological rhythms and sleep disorders, a hands-on empirical investigation was not possible. Instead, students collected data by maintaining a sleep diary in the week prior to the teaching experiments, and so these data, along with the whole class data and a number of provided data sets, provided the evidence for the investigations.

Data Scoring and Analysis

Pretest and Posttest Scoring. To score the constructed-response items we created a set of levels representing increasingly sophisticated ways of reasoning with scientific models of sleep behavior. The process of developing these levels, as well as the initial notions of the levels themselves, was modeled after the process of developing levels for a learning progression described by Chen, Mohan, and Anderson (2008). One project researcher, along with three external researchers, worked to develop levels by ranking a sample of student responses from least to most sophisticated, grouping similar responses, then characterizing and developing categories for the groups of responses. The resulting rubric allowed the full range of students' responses to be scored along a continuum from informal/non-scientific ideas about sleep, to reasoning constrained by scientific principles with models of sleep across scales. In between these extremes were responses in which students gave exclusively organismal-level accounts (i.e., focused on visible behavior) and those in which students recognized physiological control of sleep behavior, but could not describe the physiological mechanism. After development of the rubric, a blinded sample of students' pretests and posttests were scored by the group, and the extent of scoring agreement between raters was evaluated and discussed. Minor changes were made to the levels based on these discussions, and rules for scoring certain types of responses were developed. Table 5 shows the five levels that were used to score the constructed response items, along with common errors and exemplar responses at each level for one test item. Each level was further split into "High" and "Low" levels to allow greater resolution and distinction between responses. As such, each response received a reasoning score between 0 and 10.

Since four raters (one internal and three external researchers) each scored one quarter of the total set of pretests and posttests, inter-rater reliability was calculated to test for consistency in scoring between raters. A sample of 10 percent of the tests was scored by all four raters, and inter-rater reliability was calculated using the intraclass correlation coefficient. Analysis of the commonly scored items showed no significant differences between raters [$F(1, 47) = 0.033$, $p = 0.992$] with an intraclass correlation coefficient of 0.783 (two-way mixed effects model, single measures, absolute agreement). Interpretation of the intraclass correlation coefficient is similar to that of Cohen's Kappa, i.e., 0.40 to 0.59 is moderate inter-rater reliability, 0.60 to 0.79 substantial, and 0.80 outstanding (Landis & Koch, 1977). Throughout all scoring, researchers were blind to both the treatment group and whether the test was a pretest or posttest.

Hierarchical Regression. Since existing class or school structure (nesting) was not a factor in this design, multi-level modeling was not necessary. Instead we used hierarchical, ordinary least squares regression to address the questions posed in this study. We selected the order for inclusion of predictor variables so we would account for the largest, most obvious sources of variation in the outcome variable, Y (student posttest score), first. The first factor in the model was a student's pretest score. Pretest scores typically account for a high degree of variation in posttest scores (Schochet, 2005). By adding group to the model after pretest scores, we can determine the extent to which group assignment predicted variation in posttest scores, above and beyond variation already accounted for by the pretest. Thus, our assessment of group effect is more conservative than if the order of variables were reversed.

Next, we sought to determine if the benefits of group assignment were equitable across student demographic groups. In other words, we wanted to determine if student demographic variables accounted for variation in posttest scores above and beyond variation accounted for by pretest score and group assignment. If inquiry-based instruction is equitable and commonplace instruction may or may not be equitable, we would predict *no* variation in posttest scores above and beyond that accounted for by pretest and group. If we had reversed the order and added demographic variables to the model *before* either pretest or group, we would not be able to assess if variation was pre-existing, or if group assignment mitigated any pre-existing differences. Thus, we added demographic variables to the model after pretest and group, in order of their theoretical significance. Rothstein (2004) identified socioeconomic status as more important than either race/ethnicity or gender in its ability to predict student achievement. Therefore, we added FRL status as the first demographic variable in the model. Hanson (1996) and Muller, Stage, & Kinzie (2001) identified race/ethnicity as more significant predictors of science achievement than gender. Thus, race/ethnicity was the second demographic variable added to the model, followed by gender. The following model represents the final model tested:

Step 5 Model

$$\hat{Y}_{posttest} = b_o + b_1 X_{pretest} + b_2 X_{group} + b_3 X_{lunch} + b_4 X_{race} + b_5 X_{gender}$$

For Steps 2–5 of the regression, we calculated an F test of change. Each test was conducted at $\alpha = .05$. Model assumptions, including normality of residuals, homogeneity of variances, the presence of a linear relationship between the covariate (pretest) and Y (posttest), and homogeneity of regression, were met. Students were randomly assigned to either the commonplace or inquiry groups; thus, independence of residuals is likely. Furthermore, no significant correlation existed between the pretest and group.

Interview Scoring. As a framework for scoring the interviews, we began with the modification of Toulmin's argumentation model developed by McNeill et al. (2006). Students' explanations were scored according to the quality of their claim ("an assertion or conclusion that answers the original question"), evidence ("scientific data that supports the claim"), and reasoning ("a justification that shows why the data count as evidence to support the claim"). A sample of students' interviews was scored by the same four researchers as the pretests and posttests, and the extent of scoring agreement between raters was evaluated and discussed. Minor changes were made to the rubric based on these discussions, and rules for scoring certain types of responses were developed. Table 6 shows the final rubric used for interview scoring. The interviews were divided between the four scorers, and inter-rater reliability was calculated from a commonly scored random sample of six interviews. The intraclass correlation coefficient (two-way mixed effects model, single measures, absolute agreement) for the inter-rater reliability was 0.872.

Results

Total Test Scores

Students in the inquiry-based group had significantly higher posttest scores than students in the commonplace group [$F(1,55) = 4.570$, $p < 0.05$], controlling for variance in the students' pretest scores. The effect size (Cohen's d) for this difference was 0.47 (standard deviation units). We can also look at this finding in terms of the regression model shown earlier, where adding the group assignment to the model explains significantly more of the variance in posttest scores (44.3%) than pretest alone (39.7%):

$$\hat{Y}_{posttest} = \underbrace{b_o + b_1 X_{pretest}} + b_2 X_{group}$$

39.7%, $p<0.001$

44.3%, $p<0.05$

Figure 1 shows the different slopes of the pretest-posttest regression lines for each group.

Level 5 Understanding

Of the five levels used to score the constructed response items (Table 5), Level 5 (model-based accounts connected across scales) represents the type of reasoning that is a desirable goal of secondary science education (Chen et al., 2008). That is, across most (and perhaps all) science content, in order to reason scientifically, students must traverse systems across scales; keep track of matter, energy, and/or information; and connect the causes and effects of multiple processes (Wilson et al., 2006). With respect to reasoning about sleep behavior, a student with a Level 5 understanding is, for example, able to reason across physiological, organismal, and environmental systems; trace information from light cues through physiological systems; and connect processes involving light/dark cycles and hormonal signaling to account for observed behaviors. As such, we next examine how the achievement of Level 5 reasoning differed between the students in the commonplace and inquiry-based groups. Students in the inquiry-based group gave a significantly higher fraction of responses at Level 5 than students in the commonplace group, [$F(1,56) = 4.537$, $p < 0.05$], controlling for variance in the students' pretest scores. The effect size (Cohen's d) for this difference was 0.68. Figure 2 shows the effects of the two instructional units on the frequency of Level 5 accounts.

Achievement across Student Demographic Variables

In the calculation of F-change statistics for the hierarchical regression, only group assignment contributed to the model above and beyond pretest score. FRL status, race/ethnicity, and gender did not account for variation in posttest scores above and beyond other factors. Table 7 summarizes these data. Pretest score accounted for 39.7% of the variance [$F(1,58) = 36.88$, $p < 0.001$]. The addition of group assignment to the regression model significantly increased the variance explained. Pretest score and group assignment together accounted for 44.3% of the variance, $F(2,58) = 21.90$, $p < 0.001$; F-change $(1,55) = 4.54$, $p < 0.05$. In Steps 3–5, the addition of FRL, race/ethnicity, and gender did not significantly contribute to the variance explained at the 0.05 level.

To further examine differential performance as a function of FRL status, race/ethnicity, and gender in each of the two treatment groups, we examined scores on the pretest and posttest using independent t-tests. There were no significant differences by FRL status or gender on either the pretest or posttest in either group. As shown in Figure 3, the only significant difference in scores between white and non-white students was on the posttest for the students in the commonplace unit [$t(26) = 2.330$, $p = 0.028$]. That is, while there were no significant differences in the pretest scores of white and non-white students in either group, the commonplace unit resulted in significantly lower posttest scores for non-whites, yet no significant difference by race was found in the posttest scores of students in the inquiry-based group [$t(28) = 1.780$, $p = 0.086$]. That said, our sample of students within a single treatment was small and unbalanced (e.g., 23 white, 7 non-white). As a result, our study did not have the statistical power to detect within-treatment effect sizes for white and non-white students below 1.0. To further investigate the differences between inquiry and commonplace instruction, we calculated the effect sizes of the achievement gap for white and non-white students on both the pretest and posttest. While the effect size for a gap on the pretest was comparable for both groups (0.59 for commonplace and 0.64 for inquiry), the effect size on the posttest for the commonplace group (1.07) was much larger than that for the inquiry group (0.77). At a minimum, we can state that the teaching of science as inquiry mitigated the expansion of gaps that may have been present at an undetectable level in this study.

Lastly, we considered normalized gain scores (the ratio of actual gain to possible gain from pretest to posttest). Students in the inquiry group tended to show medium normalized gains (Hake, 1998), while students in the commonplace group showed low to medium gains. Furthermore, the differential in normalized gain scores between white and non-white students and male and female students was smaller in the inquiry group than it was in the commonplace group. The differential between FRL/no FRL was larger for the inquiry groups than for the commonplace group. Further investigation is needed to determine if inquiry instruction is more effective for students from high socioeconomic backgrounds. These data are summarized in Table 8.

Interviews and Argumentation

Analysis of the argumentation scores from the standardized interviews showed that students in the inquiry group had significantly higher scores for claims [$F(1,54) = 4.253$, $p < 0.05$], evidence [$F(1,54) = 9.794$, $p < 0.01$], and reasoning [$F(1,54) = 5.051$, $p < 0.05$] than students in the commonplace group. The effect sizes (Cohen's d) for each difference were 0.58, 0.74, and 0.59 respectively. We make our claims around argumentation with some caution. Although the random assignment process should (theoretically) create two groups of similar mean pretest, not having a specific argumentation covariate in the posttest argumentation model limits the precision of the treatment effect estimate. Figure 4 shows the effects of the two instructional units on students' construction and critique of explanations. Table 9 summaries the statistics from the RTOP, CLES, argumentation, reasoning and total test pre- and posttests.

Discussion and Conclusions

Using scientifically-based research methods required to establish causality, this study found that students receiving inquiry-based instruction reached significantly higher levels of achievement than students experiencing commonplace instruction. The superior effectiveness of the inquiry-based instruction was consistent across a range of learning goals (knowledge, scientific reasoning, and argumentation) and time frames (immediately following the instruction and four weeks later). This study therefore contributes to the growing body of evidence demonstrating the effectiveness of inquiry-based instruction and supports the advocacy for inquiry-based instruction stated in national and international science education reform documents (AAAS, 1993, 2000; NRC, 1996, 2000; Osborne & Dillon, 2008; Australian Education Council, 1994; Tomorrow 98, 1992; Ministry of Education, 1999). Further, findings from this study directly challenge the claims of Kirshner et al. (2006) made in response to the findings by Klahr and colleagues (Chen & Klahr, 1999; Klahr & Nigam, 2004).

Despite the long standing call for science for all (AAAS, 1993; Committee on Science, Engineering, and Public Policy [COSEPUP], 2007; NRC, 1996), achievement gaps by gender, race/ethnicity, and socioeconomic status remain in the U.S. (Clewell & Campbell, 2002). Further, learning science as inquiry may be more accessible or beneficial for some students than others (Lee, 1997; Von Secker, 2002; Barton, 2003). In this study, the results of the hierarchical regression demonstrated that race, gender, and FRL status (as a proxy for socioeconomic status) did not account for significant variation in posttest scores above and beyond pretest score and group assignment. That is, the effectiveness of the inquiry-based instruction was consistent across these variables. Examination of achievement by race in both the pretest and posttest in each treatment group revealed no significant differences by race on the pretest in either group and no significant differences by race on the posttest for the inquiry-based group; however there were significant achievement gaps in the posttest score in the commonplace group. Based on our power limitations, we can state that at a minimum, commonplace science instruction resulted in widened achievement gaps by race, whereas the inquiry-based instruction mitigated the expansion of existing gaps. These findings are consistent with those of Lynch et al. (2005), who found that students receiving inquiry-based instruction outperformed students in comparison groups, regardless of ethnicity, socioeconomic status, gender, and ESOL status, and speak to the appropriateness of inquiry and the BSCS 5Es for meeting the need of science for all.

The effect sizes in this study (total test = 0.47, scientific reasoning = 0.68, average argumentation = 0.64) are comparable to the findings from other studies recently reported in this journal. For example Geier et al. (2008) conducted a quasi-experimental, scale-up study on the effectiveness of project-based inquiry science units (supported by professional development and learning technologies) involving approximately 5000 7th and 8th grade students. The effect size (from state standardized tests) for the first cohort of teachers (as compared to business as usual teaching) was 0.44, which decreased only slightly to 0.37 during the second cohort and significant scaling of the intervention. Schroeder et al. (2007) conducted a meta-analysis on the effectiveness of various instructional approaches on student achievement, as reported from experimental and quasi-experimental studies conducted in the US. The study found an average effect size for “Inquiry Strategies” of 0.65, with the inquiry-driven approaches not being mutually exclusive from strategies with even higher effect sizes (e.g., engaging students' interest via context, effect size = 1.48; and collaborative learning strategies, effect size = 0.96). Taraban et al. (2007), in a study comparing instruction across six classrooms and 408 students, found effect sizes favoring an inquiry-based approach over traditional instruction of 0.32 (knowledge), 0.09 (critical thinking), and 0.30 (process skills). The low effect size for critical thinking in the Taraban et al. study (as compared to the high scientific reasoning effects in this study) may well be due to their use of simple multiple-choice items, rather than detailed analysis of students' accounts. Since our methodological approach was in part driven by the evidence-based reform movement, we look to see how our effect sizes compare to those from studies accepted into nationally recognized effectiveness databases. The Best Evidence Encyclopedia (BEE), a resource "intended to provide easily accessible, scientifically viable summaries of the evidence base for educational programs" found average effect sizes of 0.06 from 77 studies examining curriculum interventions; 0.11 from 130 studies on computer assisted instruction (CAI); 0.27 from 100 studies that looked at the effectiveness of instructional approaches; and 0.26 from studies that combined curriculum interventions/CAI and instructional approaches (Slavin, Lake & Davis, 2009). As such, our main effect size of the treatment, 0.47, is similar to those from other (larger) studies of instructional approaches in the BEE, while the effects found from the reasoning and argumentation measures are particularly high.

We conclude by considering why teaching science as inquiry was more effective than commonplace teaching for the learning goals measured in this study. In their review on learning in *How People Learn*, Bransford et al., (1999) describe a number of major research findings around which there is broad consensus from researchers across disciplines and content areas.

Each of these findings maps directly on explicit components of both inquiry and the BSCS 5E Instructional Model. From soliciting and building on students' prior understandings, to emphasizing deep understanding, the importance of metacognition, and the social nature of learning, both inquiry-based instruction and the 5Es mirror these findings. Both involve investigations that begin with what the student already knows; that engage students in learning content as well as how to organize and reason about the content; activities in which students control, reflect upon, and evaluate their learning; and that scaffold students working together and with the teacher to discuss evidence and connect their findings with scientific explanations. The connections between the Bransford et al. findings and both inquiry-based instruction and the 5Es are of course not coincidental, since each instructional framework was developed in response to much of the same research and evidence synthesized in *How People Learn*. As a reflection of this synthesis, the achievement measures in this study emphasize the construction of deep understanding that facilitates the retrieval and application of ideas as well as the development and construction of evidence-based arguments. Subsequently, students in the inquiry-based treatment group preformed better. On the other hand, commonplace science teaching is largely focused on a knowledge transmission model with a much narrower set of student learning goals and students receiving this treatment did not perform as well on the achievement measures.

Given the multiple and disparate definitions of inquiry across and between researchers and practitioners (Minstrell, 2000; Barman, 2002; Lederman, 2003), to examine the effectiveness of an unclearly specified enactment of inquiry is probably not particularly helpful. However, in this study we operationalized inquiry-based teaching and learning via the BSCS 5Es – an instructional model grounded in social constructivism that represents a purposeful organization and sequence of inquiry teaching strategies. This raises further questions, since interpretations and implementation of the 5Es can be just as inconsistent as that of inquiry. However, we contend that providing teachers with well-designed curriculum materials removes many of the ambiguities associated with inquiry, and such an instructional model–guided approach to teaching and learning is supported by significant national reports (Bransford et al., 1999). This study also has implications for the development and use of curriculum materials. Since teacher effects were removed and the students were randomly assigned to treatments, we can be more confident in attributing the effects found in this study to the curriculum materials and their embedded strategies. As such, our findings reinforce the hypothesis that inquiry-based teaching can be supported by research-based curriculum materials (Brown & Edelson, 2003; Davis & Krajcik, 2005; Remillard, 2005).

Blanchard et al. (2008) describe how many teachers perceive teaching for accountability and teaching via inquiry to be incompatible, yet the findings presented here substantiate the claim that this is indeed a false and unnecessary dichotomy. Because students in the inquiry-based group outperformed students receiving commonplace instruction on each of the knowledge, scientific reasoning, and argumentation measures, this study provides evidence that teachers need not compromise the quality of their teaching (Shaver et al., 2007; Southerland et al., 2007) to see increases in student achievement. It is especially worth noting that the retention of ideas was stronger for the students in the inquiry-based experience (as measured by the delayed argumentation posttest with a general pretest covariate) because high-stakes testing typically occurs once a year and is therefore dependent on students retaining ideas for long periods of time. Because the learning goals measured in this study align with those described in the NSES (NRC, 1996, 2000), and as state standards and tests continue to converge on the national standards, we suggest that inquiry-based teaching and learning is not discordant with the pressures of accountability and high-stakes testing.

In investigating this approach to science teaching and learning, we concur with Hmelo-Silver, Duncan, and Chinn's statement (2007): "Does it work? is the wrong question. The more important questions to ask are under what circumstances do these guided inquiry approaches work, what are the kinds of outcomes for which they are effective, what kinds of valued practices do they promote, and what kinds of support and scaffolding are needed for different populations and learning goals" (p105). This study makes a significant step towards addressing those more nuanced questions by examining the effects of inquiry-based instruction on multiple, relevant learning goals (knowledge, reasoning, and argumentation), and by looking at those effects across different populations. The survey and interview data from Horizon Research, Inc. (Weiss et al., 2003; Hudson et al., 2002) highlight the disparity between the central position inquiry holds in science education reform and its placement on the periphery of practice in science classrooms. While there are many legitimate barriers to inquiry-based teaching and enacting inquiry-based curriculum materials, we hope that this and complementary studies help minimize the constraints presented by certain political, cultural, and even technical dilemmas (Anderson, 2002). By meeting the standards of evidence required in a climate of accountability and evidence-based reform, this work provides support for the continued transition of inquiry-based teaching and learning from theory and advocacy to practice and policy.

Limitations of the Study

There are a number of limitations to this study that should be noted. As described previously, statistical conclusion validity in the argumentation analysis was limited by not having a specific argumentation pretest covariate in the model, and instead a knowledge/reasoning pretest value for each student was used because a correlation was expected. Our claims about retention are similarly tempered (no argumentation pretest or no knowledge reasoning retention measure). Other limitations of this study include the small sample size (58 students) and the short length of the intervention (10 hours of instruction, 4 hours of testing), yet the fact that we found significant and consistent differences despite these limitations speaks to the strength of the effect. The laboratory-based randomized control design with a controlled teacher variable led to a study with high internal validity, but with the external validity being somewhat compromised by the lack of a random or stratified sample, and by the clinical/non-school–based setting for the instruction. However, in another sense, external validity was increased by our comparison to commonplace instruction (operationally defined through the use of data from large-scale teacher surveys; Weiss et al., 2003; Hudson et al., 2002) instead of merely didactic or direct instruction, which are the commonly used counterfactuals in studies of this kind (Klahr & Nigam, 2004; Lederman et al., 2008). Despite the teacher in this study having many years experience teaching both traditional and inquiry-based materials, he is undoubtedly more of an advocate of an inquiry-based approach. However, we believe the benefits of controlling variables by having the same teacher in both sections outweighed the potential bias created by a teacher being more comfortable in one approach than the other, and findings such as the comparable levels of student engagement shown in Table 4 suggest that the treatments were not strongly teacher-biased.

Researchers across disciplines have always faced the dilemmas of balance and compromise between internal and external validity, yet such questions take on increased significance when one approach to research becomes overwhelmingly advocated by those holding the keys to policy. Some critics of the rhetoric of randomized control trials (and other methodological approaches characteristic of evidence-based reform) argue that teaching and learning are too context-bound to allow one to generalize effectiveness to other settings (Chatterji, 2008; Green & Skukauskaitė, 2008). The question therefore becomes, to what extent can we generalize the effects found in this study? Since any single study cannot address all questions of variable impacts across every possible audience (e.g., students, teachers, schools, communities, program implementation; Briggs, 2008) generalizability follows from placing our findings in the context of other research studies, such as those described previously in the reviews by Hmelo-Silver et al. (2007) and Colburn (2008). While this study is situated within the boundaries of the United States, its findings inform an international context where inquiry-based instruction is valued. Indeed, inquiry is one of only a handful of themes that are common in K-12 science standards and curricula around the world (Abd-El-Khalick et al., 2004). While many questions still exist, this study complements the existing body of evidence on the effectiveness of inquiry-based teaching and learning, and extends that evidence to encompassing a broader range of philosophical and methodological traditions.

Author Note

Christopher D. Wilson, Center for Research and Evaluation, BSCS; Joseph A. Taylor, Center for Research and Evaluation, BSCS; Susan M. Kowalski, Center for Research and Evaluation, BSCS; Janet Carlson, BSCS.

This research was supported by the Office of Science Education at the National Institutes of Health. The authors would like to thank the external researchers, Martha Alvarez, Sharon Harter, and Sandra Smith, for scoring tests and interviews; David Pinkerton for his role as teacher and his work on module development; as well as Mark Bloom and April Gardner for their assistance with interviews and module development. We also thank Nancy Landes and Brett Merritt for their comments on the manuscript.

Correspondance concerning this article should be addressed to Christopher Wilson, BSCS Center for Research & Evaluation, 5415 Mark Dabling Blvd., Colorado Springs, CO 80918. E-mail: cwilson@bscs.org

References

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H.-L. (2004). Inquiry in science education: International perspectives. *Science Education, 88,* 397–419.

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy.* New York: Oxford University Press.

American Association for the Advancement of Science (AAAS). (2000). *Designs for science literacy.* New York: Oxford University Press.

Anderson, C. W. (2003). *Teaching science for motivation and understanding.* Unpublished manuscript. Retrieved August 2008, from: https://www.msu.edu/~andya/TEScience/Assets/Files/TSMU.pdf

Anderson, R. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13,* 1–2.

Australian Education Council (1994). A national statement on science for all Australian schools: A joint project of the states, territories, and commonwealth of Australia initiated by the Australian Education Council (AEC). Carton, Vic: Curriculum Corporation.

Barman, C. (2002). How do you define inquiry? *Science & Children. 40*(2): 8–9.

Barton, A. C. (2003). *Teaching science for social justice.* New York: Teachers College Press.

Blanchard, M. R., Annetta, L. A., & Southerland, S. A. (2008). *Investigating the effectiveness of inquiry-based versus traditional science teaching methods in middle and high school laboratory settings.* Paper presented at the annual conference of the National Association for Research in Science Teaching, Baltimore, MA.

Bransford, J., Brown, A., & Cocking, R. (Eds.) (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Briggs, D. C. (2008). Synthesizing causal inferences. *Educational Researcher, 37,* 15–22.

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Brown, M., & Edelson, D. (2003). *Teaching as design: Can we better understand the ways in which teachers use materials so we can better design materials to support changes in practice?* Research Report, Center for Learning Technologies in Urban Schools (Northwestern University).

BSCS. (2003). *Sleep, sleep disorders, and biological rhythms.* Bethesda, MA: National Institutes of Health (NIH) (NIH Publication No. 04-4989).

Bybee, R. W. (1997). *Achieving scientific literacy: From purposes to practices.* Portsmouth, NH: Heinemann Educational Books.

Bybee, R. W., Carlson Powell, J., & Trowbridge, L. W. (2007). *Teaching secondary school science: strategies for developing scientific literacy.* Boston: Prentice Hall.

Bybee, R. W., & Landes, N. M. (1990). Science for life and living. *American Biology Teacher, 52*(2), 92–98.

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Carlson, J., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins, effectiveness, and applications.* Unpublished white paper. Retrieved August 2008, from http://www.bscs.org/pdf/5EFull Report.pdf

Carlson, R. A., Lundy, D. H., & Schneider, W. (1992). Strategy guidance and memory aiding in learning a problem-solving skill. *Human Factors, 34,* 129–145.

Cartier, J., Rudolph, J., & Stewart, J. (2001). *The nature and structure of scientific models.* Madison, WI: The National Center for Improving Student Learning and Achievement in Mathematics and Science (NCISLA).

Chatterji, M. (2008). Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher, 37*, 23–26.

Chen, J., Mohan, L., & Anderson, C. W. (2008). *Developing a K-12 learning progression for carbon cycling in socio-ecological systems.* Paper presented at the 2008 annual meeting of the National Association for Research in Science Teaching, Baltimore MD.

Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Clewell, B. C., & Campbell, P. B. (2002). Taking stock: Where we've been, where we are, where we're going. *Journal of Women and Minorities in Science and Engineering, 8,* 255–284.

Colburn, A. (2000). An inquiry primer. *Science Scope,* 23(6), 42–44.

Colburn, A. (2008). *What teacher educators need to know about inquiry-based instruction.* Retrieved February 22, 2008, from http://www.csulb.edu/~acolburn/AETS.htm

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.

Committee on Science, Engineering, and Public Policy (COSEPUP), National Academy of Sciences, National Academy of Engineering, and Institutes of Health.. (2007). *Rising* *above the gathering storm: Energizing and employing America for a brighter economic future.* Washington, DC: National Academy Press.

Confrey, J. (2007). Comparing and contrasting the National Research Council report on evaluating curricular effectiveness with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis, 28*(3), 195–213.

Crawford, B. A. (2007). Learning to teach science as inquiry in the rough and tumble of practice. *Journal of Research in Science Teaching 44*(4), 613-642.

Cuban, L. (1988). Constancy and change in schools (1880s to the present). In P. W. Jackson (Ed.), *Contributing to educational change* (pp. 85–105). Berkley, CA: McCutchan.

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher, 34*(3), 3–14.

Dean, D., & Kuhn, D. (2006). Direct instruction vs. discovery: The long view. *Science Education, 91,* 384–397.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312.

Gallagher, J. J. (1989). Research on secondary school science practices, knowledge and beliefs: a basis for restructuring. In M. Matayas, K. Tobin, & B. Fraser (Eds.), *Looking into windows: qualitative research in science education.* Washington, DC: American Association for the Advancement of Science.

Geire, R. (1999). Using models to represent reality. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery.* New York: Kluwer Academic/Plenum Publishers.

Geier, R., Blumenfeld, P., Marx, R., Krajcik, J., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry based science curriculum in the context of urban reform. *Journal of Research in Science Teaching, 45* (8) 922-939.

Gilbert, J. K., Boulter, C. J., & Rutherford, M. (1998). Models in explanations, part 1: Horses for courses? *International Journal of Science Education, 20*(1), 83–97.

Green, J. L., & Skukauskaite, A. (2008). Becoming critical readers: Issues in transparency, representation, and warranting of claims. *Educational Researcher, 37,* 30–40.

Golan, R., Kyza, E. A., Reiser, B. J., & Edelson, D. C. (2002, April). *Scaffolding the task of analyzing animal behavior with the Animal Landlord software.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hake, R. R. (1998) Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics. 66,* 64–74.

Hall, D. A., & McCurdy, D. W. (1990). A comparison of a Biological Sciences Curriculum Study (BSCS) laboratory and a traditional laboratory on student achievement at two private liberal arts colleges. *Journal of Research in Science Teaching, 27,* 625–636.

Hanson, S. L. (1996). *Lost talent: Women in the sciences.* Philadelphia, PA: Temple University Press.

Hardiman, P., Pollatsek, A., & Weil, A. (1986). Learning to understand the balance beam. *Cognition and Instruction, 3,* 1–30.

Hickey, D. T., Kindfeld, A. C. H., Horwitz, P., & Christie, M. A. (1999). Advancing educational theory by enhancing practice in a technologysupported genetics learning environment. *Journal of Education, 181,* 25– 55.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research.* MDRC. Retrieved August 11, 2008 from http://www.mdrc.org/publications/459/full.pdf.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42,* 99–107.

Hudson, S. B., McMahon, K. C., & Overstreet, C. M. (2002). *The 2000 national survey of science and mathematics education: Compendium of tables.* Chapel Hill, NC: Horizon Research.

Jackson, S., Stratford, S. J., Krajcik, J. S., & Soloway, E. (1996). Making system dynamics modeling accessible to pre-college science students. *Interactive Learning Environments, 4,* 233–257.

Kirschner, P.A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry based teaching. *Educational Psychologist, 41,* 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15,* 661–667.

Krajcik, J. S., Czerniak, C., & Berger, C. (1999). *Teaching children science: A project-based approach.* Boston: McGraw-Hill.

Landis, J. R., & Koch, G. (1977). *The measurement of observer agreement for categorical data. Biometrics, 33*(1), 159–174.

Lawrenz, F., Huffman, D., & Gravely, A. (2007). Impact of the collaboratives for excellence in teacher preparation program. *Journal of Research in Science Teaching, 44*(9), 1348-1369.

Lawson, A. E. (1995). *Science teaching and the development of thinking.* Belmont, CA: Wadsworth Publishing.

Lederman, N. (2003). Letters: Learning about inquiry. *Science & Children, 40*(8), 9.

Lederman, N. (2004). Scientific inquiry and science education reform in the United States. In F. Abd-El-Khalick, S. Bougaoude, N. Lederman, A. Mamok-Naaman, Hopstein, M. Nioz, D. Treagrest, & H. Tusan (Eds.), Inquiry in science education: International perspective (pp. 402–404). *Science Education, 88,* 397–419.

Lederman, N., Lederman, J., & Wickman, P.-O. (2008). *An international, systematic investigation of the relative effects of inquiry and direct instruction: A replication study.* Paper presented at the annual conference of the National Association for Research in Science Teaching, Baltimore, MA.

Lee, O. (1997). Scientific literacy for all: What is it, and how can we achieve it? *Journal of Research in Science Teaching, 34,* 219–222.

Leonard, W. H. (1983). An experimental study of a BSCS-style laboratory approach for university general biology. *Journal of Research in Science Teaching, 20,* 807–813.

Leonard, W. H., Cavana, G. R., & Lowery, L. F. (1981). An experimental test of an extended discretion approach for high school biology laboratory investigations. *Journal of Research in Science Teaching, 18,* 497–504.

Lewis, S. E. & Lewis, J. E. (2008). Seeking effectiveness and equity in a large college chemisty course: an HLM investigation of a peer-led guided inquiry. *Journal of Research in Science Teaching, 45*(7), 794-811.

Lotter, C., Harwood, W. S., & Bonner, J. J. (2007). The influence of core teaching conceptions on teacher's use of inquiry teaching practices. Journal of Research in Science Teaching, 44, 1318-1347.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching, 42,* 921–946.

McGinnis, R. Parker, P., & Graeber, A. (2004). A cultural perspective of the induction of five reform-minded beginning mathematics and science teachers. *Journal of Research in Science Teaching, 41,* 720–747.

McNeill, K. L., Lizotte, D. J, Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. The Journal of the Learning Sciences, 15(2), 153–191.

Ministry of Education. (1999). Curriculum outline for "Nature science and living technology." Taipei: Ministry of Education. (In Taiwanese)

Minstrell, J. (2000). Implications for teaching and learning inquiry: A summary. In J. Minstrell & E. van Zee (Eds.), Inquiring into inquiry learning and teaching in science (pp. 471–496). Washington, DC: American Association for the Advancement of Science.

Moreno, R. (2004). Decreasing cognitive load in novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32,* 99–113.

Muller, P. A., Stage, F. K., & Kinzie, J., (2001). Science achievement growth trajectories: Understanding factors related to gender and racial-ethnic differences in precollege science achievement. *American Educational Research Journal, 38,* 981–1012.

National Research Council (NRC). (1996). *National science education standards.* Washington, DC: National Academy Press.

National Research Council (NRC). (2000). *Inquiry and the national science education standards.* Washington, DC: National Academy Press.

Osborne, J. F. & Dillon, J. (2008) Science education in Europe: Critical reflections. A Report to the Nuffield Foundation.

Osborne, J. F. (2009). Translating research into practice in the teaching of science. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Diego, CA.

Piburn, M., Sawada, D., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). Reformed Teaching Observation Protocol (RTOP) ACEPT IN-003.

Remillard, J. T. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry, 29,* 315–342.

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75*(2), 211–246.

Roehrig, G. H., & Luft, J. A. (2004). Constraints experienced by beginning secondary science teachers in implementing scientific inquiry lessons. *International Journal of Science Education, 26*(1), 3–24.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap.* Washington, DC: Teachers College Press.

Rutherford, F. J., & Ahlgren, A. (1989). *Science for all Americans.* New York: Oxford University Press.

Sawada, D., Piburn, M., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics, 102*(6), 245–253.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49,* 31–57.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13–21.

Schmidt, H. G. (1983). Problem-based learning: rationale and description. *Medical Education, 17,* 11–16.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T-Y., & Lee, Y-H. (2007). A meta-analysis of national research: effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436-1460.

Schwab, J. J. (1962). The teaching of science as enquiry. In J. J. Schwab & P. F. Brandwein (Eds.), *The teaching of science.* Cambridge, MA: Harvard University Press.

Schochet, P. A. (2005). *Statistical power for random assignment evaluations of education programs.* Princeton, NJ: Mathematica Policy Research.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16,* 475–522.

Shaver, A., Cuevas, P., Lee, O., & Avalos, M. (2007). Teachers' perceptions of policy influences on science instruction with culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching, 44*(5), 725–746.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education. Committee on Scientific Principles for Education Research. Division on Behavioral and Social Sciences and Education*. Washington, DC: National Academy Press.

Shymansky, J. A., Yore, L. D., Annetta, L. A., & Everett, S. A. (2008, January*). The impact of a five-year, externally funded, K-6 systemic reform effort on elementary school students' achievement in science.* Paper presented at the annual meeting of the Association for Science Teacher Education, St. Louis, MO.

Shymansky, J. A., Kyle, W. C., Jr., & Alport, J. M. (1983). The effects of new science curricula on student performance. *Journal of Research in Science Teaching, 20,* 387–404.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37,* 5–14.

Slavin, R. E., Lake, C. & Davis, S. (2009). *Meta-findings from the Best Evidence Encyclopedia.* Paper presented at the annual meeting of the Society for Research on Education Effectiveness, Crystal City, VA.

Southerland, S. A., Abrams, E., & Hutner, T. (2007). The accountability movement and inquiry: Must they be mutually exclusive demands? In E. Abrams, S. Southerland, & P. Silva (Eds.), *Inquiry in the science classroom: Challenges and opportunities.* Greenwich, CT: Information Age Publishing.

Taraban, R., Box, C., Myers, R., Pollard, R., & Bowen., C. W. (2007). Effects of active-learning experiences on achievement, attitudes, and behaviors in high school biology. *Journal of Research in Science Teaching, 44*(7), 960-979.

Taylor, P. C., & Fraser, B. J. (1991). *Development of an instrument for assessing constructivist learning environments.* Paper presented at the annual meeting of the American Educational Research, 27(4), 293–302.

Tobin, K., & McRobbie, C. J. (1996). *Cultural myths as constraints to the enacted curriculum.* Science Education, 80, 223-241.

Tomorrow 98. (1992). Report of the superior committee on science, mathematics and technology in Israel. Jerusalem: Ministry of Education and Culture. (English edition: 1994).

Toulmin, S. (1958). The uses of argument. Cambridge, MA: Cambridge University Press.

U.S. Department of Education. (2002). *No Child Left Behind: A desktop reference.* Washington, DC: Available from http://www.ed.gov/offices/OESE/reference.

Von Secker, C. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *The Journal of Educational Research, 95,* 151–160

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K-12 mathematics and science education in the U.S.* Chapel Hill, NC: Horizon Research.

Welch, W. W., Klopfer, L. E., & Aikenhead, G. E. (1981). The role of inquiry in science education: Analysis and recommendations. *Science Education, 65,* 33–50.

Westbrook, S. L., & Rogers, L. N. (1994). Examining the development of scientific reasoning in ninth-grade physical science students. *Journal of Research in Science Teaching, 31,* 65–76.

Whitford, B. L., & Jones, K. (2000). Kentucky lesson: How high stakes school accountability undermines a performance-based curriculum vision. In B. L. Whitford & K. Jones (Eds.), *Accountability, assessment, and teacher commitment: Lessons from Kentucky's reform efforts* (pp. 9–24). Albany, NY: State University of New York Press.

Wilson, C. D., Anderson, C. W., Heidemann, M., Merrill, J. E., Merritt, B. W., Richmond, G., Sibley, D. F., & Parker, J. M. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. *CBE Life Sciences Education, 5*(4), 323-331.

Wise, K. C., & Okey, J. R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. *Journal of Research in Science Teaching, 20*(5), 419–435.

Woodbury, S., & Gess-Newsome, J. (2002). Overcoming the paradox of change without difference: A model of change in the arena of fundamental school reform. *Educational Policy, 16*(5), 763–782.

Appendix

Example Questions from the Pretests, Posttests, and Interviews

Dichotomous Pretest and Posttest Items (Scientific Knowledge)

1. What are Circadian rhythms?

   a) Cycles that regulate our 24-hour sleep/wake cycle.

   b) The cycles between REM and NREM sleep throughout the night.

   c) The signals that let our bodies know when we have had enough sleep.

   d) The stages of the moon that regulate our sleep patterns.

True or False:

2. Sleep is a time when the body and brain shut down for rest.

3. The different stages of sleep throughout the night are EEG, EMG, and EOG.

Constructed Response Pretest and Posttest Items (Scientific Reasoning through Application of Models)

1. A person travels on a plane from Denver, CO, to London, England. London is 7 hours ahead of Denver. On the two graphs below, draw and shade in the area when the person might be asleep one day after arriving in London, and one week after arriving in London. Under each graph, explain why you shaded the area you did. Note that the times on all graphs are in Denver time.

FIGURE A1

2. Below are graphs of Ken and Annabelle's sleep patterns. Underneath each graph, describe their sleep patterns in as much detail as you can, and describe what, if anything, might be causing their sleep patterns.

FIGURE A2

Sample Interview Questions (Construction and Critique of Scientific Explanations)

Take a look at the following graph—it's a little different from the one you just saw. Take a moment to familiarize yourself with it.

FIGURE A3

Can you describe the patterns you see in the figure?

Do you see any thing else happening during the person's sleep/wake cycles?

Here's how a student explained the patterns in the graph:

- *Student response: "The person is probably living in a cave or some place with no light. Without exposure to light each day, their biological clock will not function properly."*

Can you tell us what you think of this explanation, and why you think the student may or may not be correct?

Can you tell us what you think might explain the pattern in this graph?

Possible Probes:

- You've said something about what is going on in the person's environment, what about inside their body?

- So you've described how this person might be (in a cave, experiencing jetlag, suffering from a sleep disorder, etc.), can you think of anything else that might be causing this pattern?
- Did you have any ideas that you decided were probably not correct? What were they, and why did you decide against them?

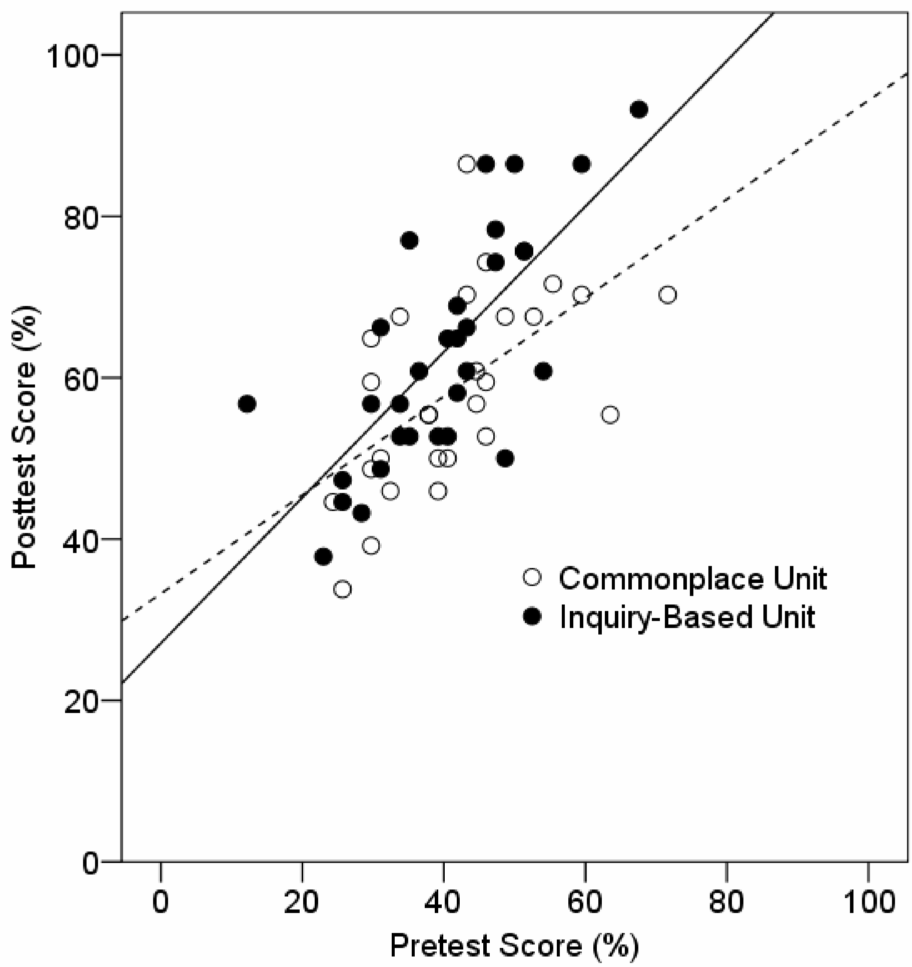


Figure 1. Pretest-Posttest bivariate distribution for the students receiving instruction from the commonplace and inquiry-based units. The slopes of the regression lines are significantly different [$F(1,55) = 4.662$, $p < 0.05$].

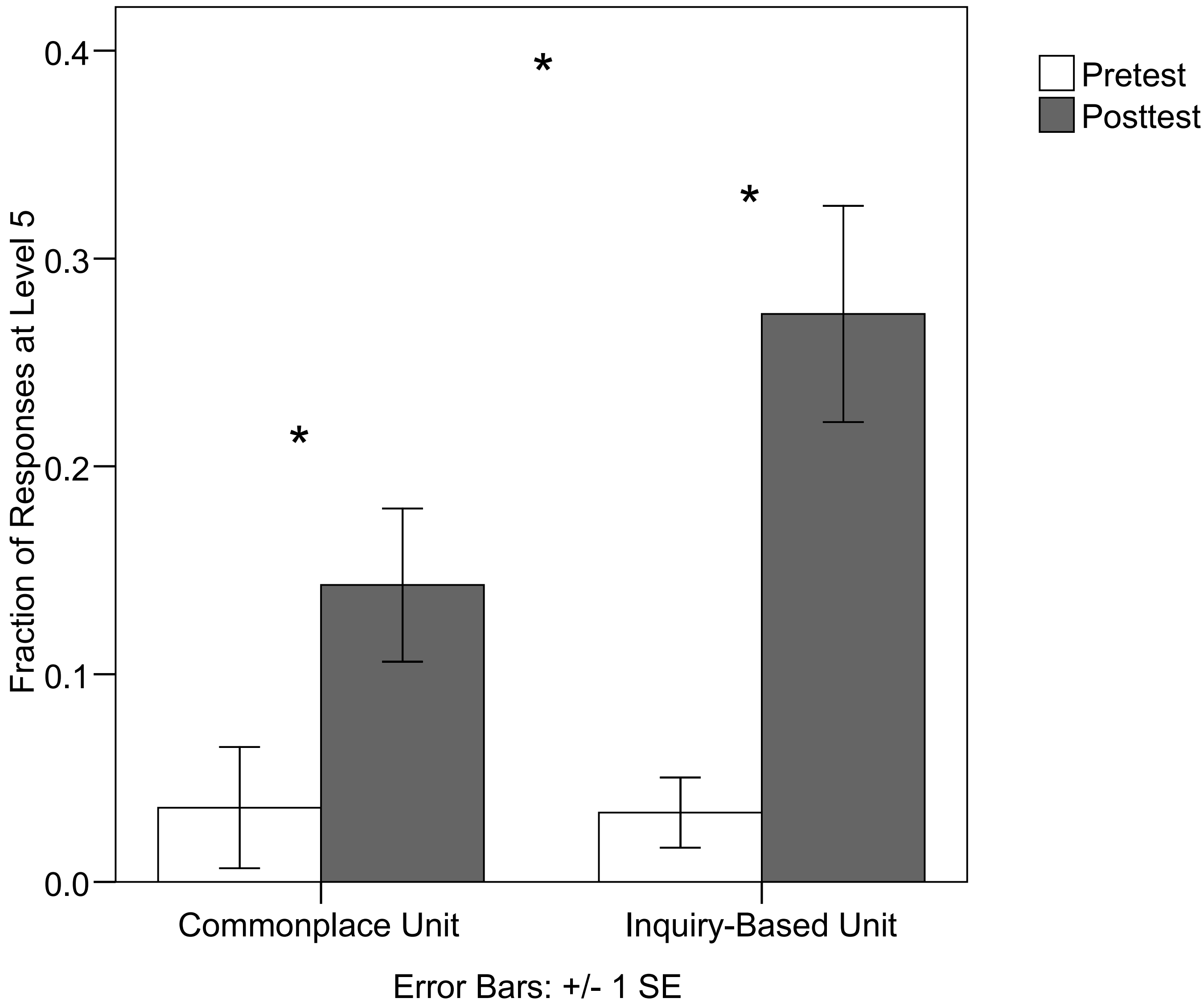


Figure 2. Significant differences [$F(1,56) = 4.537$, $p < 0.05$] in the frequency of posttest Level 5 accounts between the commonplace and inquiry-based groups. Error bars = +/- 1SE.

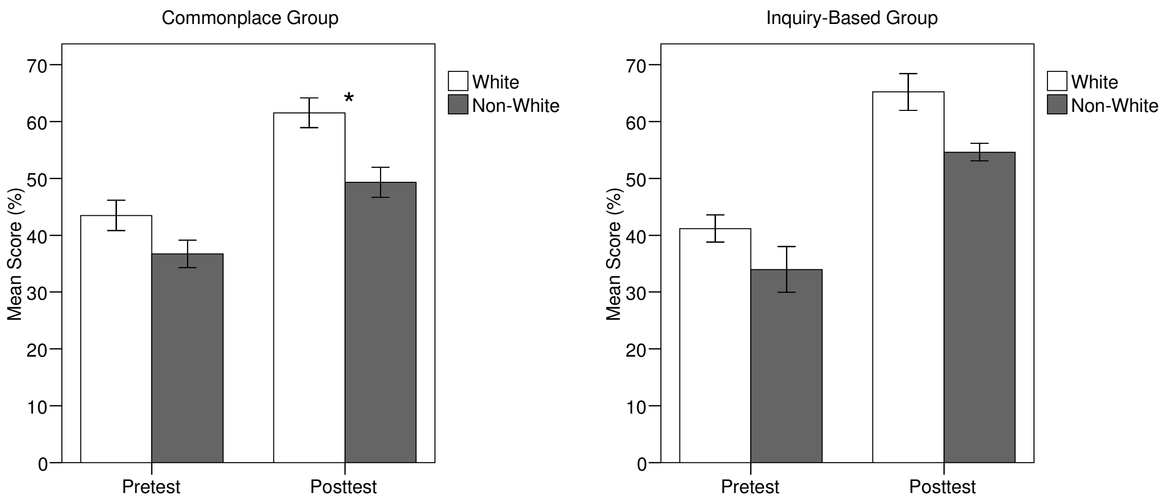
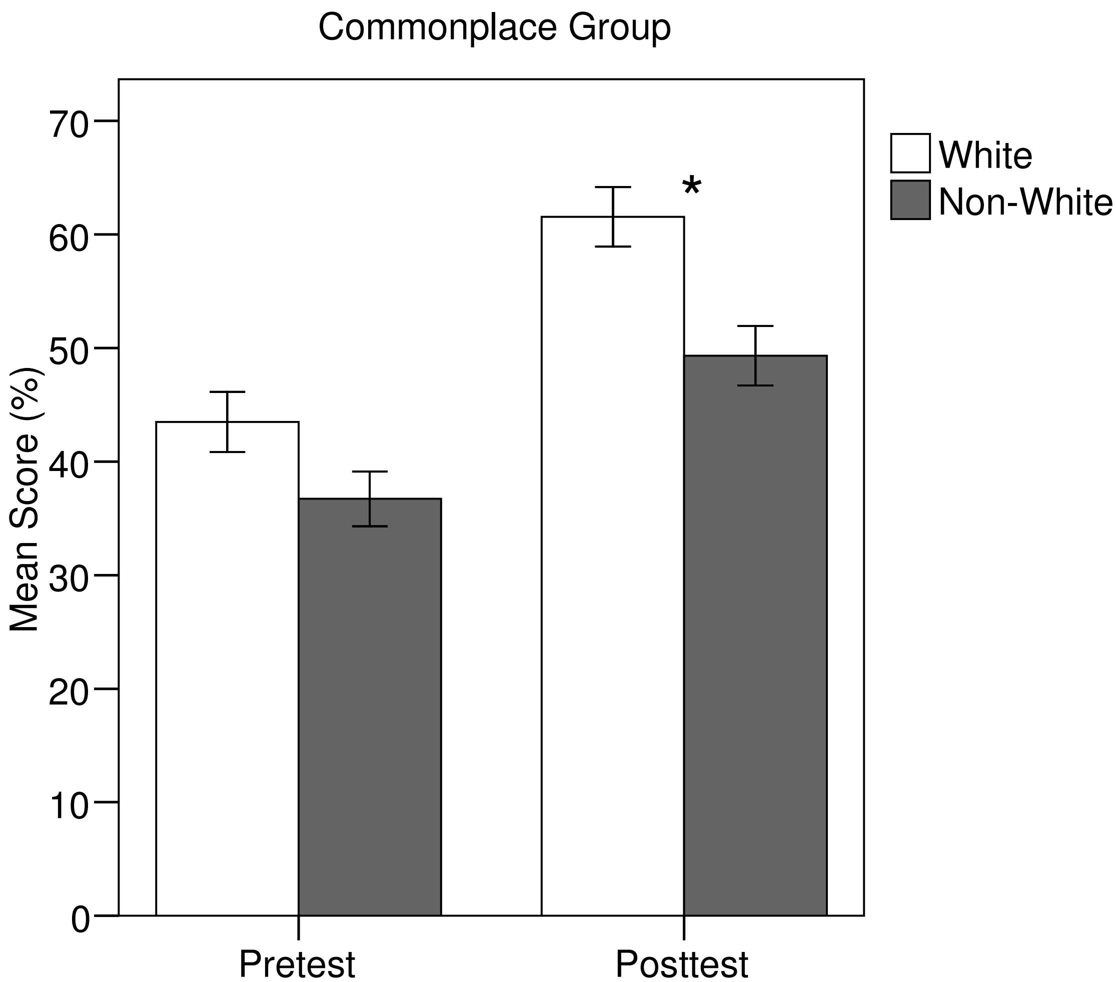


Figure 3. Differences between the pretest and posttest score of students by race in the commonplace and inquiry-based units. The only significant difference ($p<0.05$) was in the posttest scores of students in the commonplace unit [$F(1,27) = 5.530$, $p = 0.026$], indicating that the commonplace science teaching led to a significant achievement gap by race, whereas the inquiry-based instruction did not. Asterisks indicate significant differences of $p<0.05$, error bars = +/- 1SE.

(Full size graphs are below)


Commonplace Group
White
Non-White
*
Mean Score (%)
0
10
20
30
40
50
60
70
Pretest
Posttest

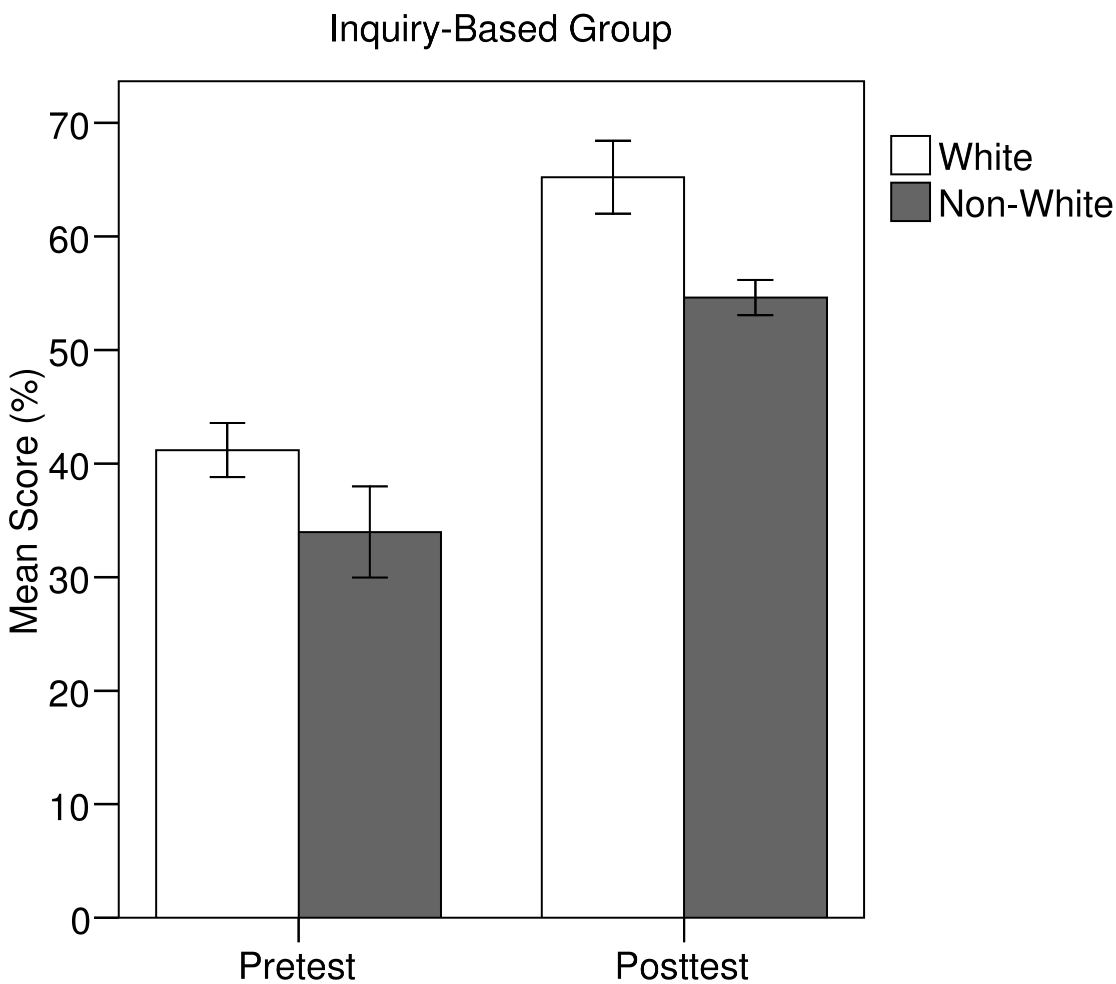

Inquiry-Based Group
Mean Score (%)
70
60
50
40
30
20
10
0
Pretest
Posttest
White
Non-White

Figure A1.

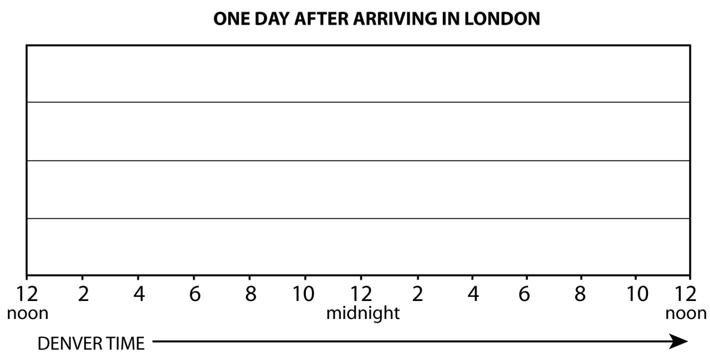

ONE DAY AFTER ARRIVING IN LONDON
12 noon
2
4
6
8
10
12 midnight
2
4
6
8
10
12 noon
DENVER TIME

Figure A2.

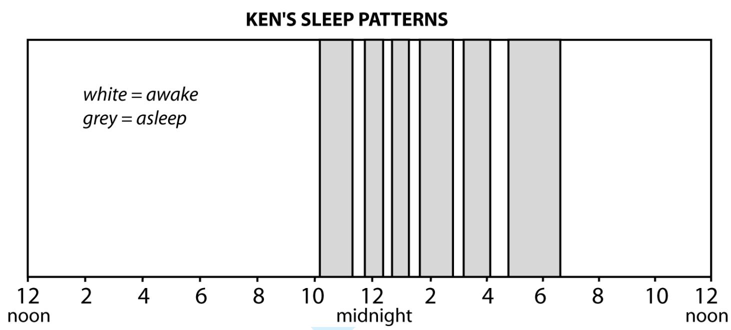

KEN'S SLEEP PATTERNS
white = awake
grey = asleep
12
noon
2
4
6
8
10
12
midnight
2
4
6
8
10
12
noon

Figure A3.

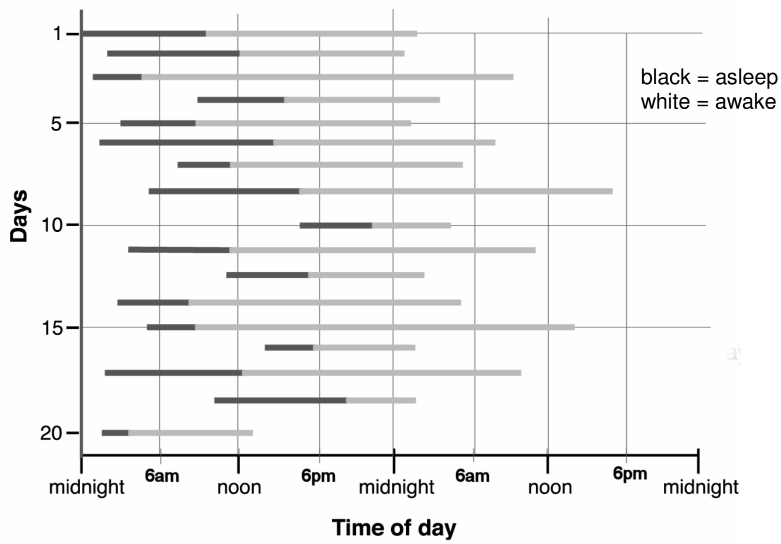

black = asleep
white = awake
Days
1
5
10
15
20
midnight
6am
noon
6pm
midnight
6am
noon
6pm
midnight
Time of day

Table 1. Alignment of the BSCS 5E model's student roles with the NSES (NRC, 2000).

The Five Essential Features of Inquiry	**The BSCS 5Es – "What the Student Does"**
Learners are engaged by scientifically oriented questions.	• Asks questions such as, "Why did this happen?" "What do I already know about this?" "What can I find out about this?" • Shows interest in the topic
Learners give priority to evidence, which allows them to develop and evaluate explanations that address scientifically oriented questions.	• Tests predictions and hypotheses • Records observations and explanations • Forms new predictions and hypotheses
Learners formulate explanations from evidence to address scientifically oriented questions.	• Uses recorded observations in explanations • Develops explanations based on data
Learners evaluate their explanations in light of alternative explanations, particularly those reflecting scientific understanding	• Forms new predictions and hypotheses • Tries alternatives and discusses them with others • Compares personal explanation with scientifically accepted explanation • Assesses own understanding
Learners communicate and justify their proposed explanations	• Explains possible solutions or answers to others • Listens critically to others' explanations • Questions others' explanations • Checks for understanding among peers

Table 2. Summary of student demographic data.

	Commonplace Unit (n=28)	Inquiry-Based Unit (n=30)
Gender	61% male, 39% female	47% male, 53% female
Race (% non-white)	21%	23%
Age (mean)	15.1	14.9
Free and Reduced Lunch	12%*	10%

*n = 26, two home-schooled students in the commonplace group did not answer this question.

Table 3. Summary description of key student activities in each classroom session.

Day	Commonplace	Inquiry
1	• record individual sleep diary data on classroom chart • complete a worksheet answering two teacher-provided questions associated with class data with step by step instructions for answering questions	• take pre-assessment on sleep misconceptions • receive instruction on writing scientific questions, then pose their own questions related to sleep diary data and write procedures for answering questions and record data from individual sleep diaries on classroom chart
2	• continue answering questions using data from day before • listen to teacher discuss patterns in data • take notes as teacher lectures on biological rhythms/cycles • read Michel Siffre cave story and answer questions about story • take notes on how to graph sleepiness data and look for patterns, then graph own data and describe patterns • participate in classroom discussion with teacher lecture about patterns • complete a worksheet answering question about a provided graph, then participate in a class discussion about responses to worksheet	• write in notebooks what they consider essential in any step by step investigation plan, share ideas in group, and revise. Then listen to ideas presented by teacher • revise previous day's procedure using newly developed essential features of investigations list, carry out experiment and prepare a claim on poster paper using explanation template; present findings from experiments to others in a "gallery walk" • receive graph template for graphing sleepiness data. Then pose questions, design experiments, and carry out experiments associated with sleepiness data; provide feedback to others on experiments
3	• take notes on stages of sleep and instruments used to measure sleep • work on astronaut scenario problem using information from notes • listen as teacher goes over astronaut scenario, providing answers to students	• read Michel Siffre cave story, construct explanations of Michel Siffre sleepiness data, share explanations with group, and revise answers • take notes as teacher lectures on key terms • analyze a new sleepiness graph to determine which sleep disorder best accounts for the graph • work on astronaut scenario problem using a claims, evidence, and reasoning template
4	• listen as teacher presents Monday morning blues and jet lag analogy problem, creates all graphs for students on the board, and explains to students why Monday morning blues is most like jetlag flying East • take notes on sleep disorders • review case studies of people with sleep disorders. Using notes and information from a reading, students attempt to diagnose each sleep disorder. Then, the teacher presents answers to students	• listen as teacher presents Monday morning blues and jet lag analogy problem. Teacher creates a sleepiness graph for Monday morning blues, then asks students to create sleepiness graphs for jet lag flying East and jet lag flying West. Students are to analyze their own graphs and use the evidence in the graphs to solve the problem • review case studies of people with sleeping disorders. Using information from a reference manual, students attempt to diagnose each sleep disorder. Students are split into groups based on their diagnoses, and participate in a class discussion defending their diagnoses
5	• listen as teacher presents common misconceptions about sleep to students (same misconceptions given to students in inquiry group as a pre-assessment on day 1)	• create their own questions relating to key concepts of the unit. Students work in pairs, one student creating a question to be answered using a verbal response, the other student creating a question to be answered using either diagrams or graphs
6	• create a crossword puzzle of key terms in order to review the concepts	• attempt to answer the questions their teammates wrote the previous day.

Table 4. Classroom differences by activity, organization, student attention, cognitive activity, and concept.

Classroom Activity	Total Minutes Commonplace	Total Minutes Inquiry
Lecture	265	145
Problem Modeling	20	0
Lecture with Discussion	15	25
Utilizing Digital Educational Media and/or Technology	5	0
Class Discussion	0	10
Writing Work	160	275
Reading Seat Work	30	50
Hands-on Activity/Materials	0	0
Small Group Discussion	70	245
Teacher/Faculty Member Interacting with Students	5	100
Administrative Tasks	15	15
Student Attention to Lesson		
Low (80% or more off task)	5	0
Medium (mixed engagement)	45	30
High (80% or more of the students engaged)	435	475
Cognitive activity		
Receipt of knowledge	435	205
Application of procedural knowledge	35	40
Knowledge representation	5	30
Knowledge construction	25	235
Concept		
What do you know about sleep/sleep misconceptions	40	20
Analysis of sleep diaries/biological rhythms	125	115
Sleep disorders, their symptoms, and case study diagnoses	130	85
Sleep Patterns and Sleepiness Scale	25	55
Writing a scientific question	0	10
Writing a scientific procedure	0	15
Analogy Problem: Monday Morning Blues and Jet Lag	20	60
Sleep Cycles, Measuring Sleep Cycles and the Astronaut Problem	85	60
Review of Big Ideas	45	55

Table 5. Description of the levels used to score the pretest and posttest constructed response items. Exemplar responses are from the item:

A person travels on a plane from Denver CO, to London, England. London is 7 hours ahead of Denver. On the two graphs below, draw and shade in the area when the person might be asleep one day after arriving in London, and one week after arriving in London. Under each graph, explain why you shaded the area you did.

Level	Description	Common Errors	Exemplar
5	Model-based accounts connected across scales	Responses may contain errors such as east/west time zone mix-ups, or details of REM-NREM cycling.	*"Despite the new light cues in London, they would still be sleeping on the Denver time because their biological clock can't reset that quickly."*
4	Appropriate but superficial connections between organismal and physiological systems	Recognizes that an internal biological clock plays a role in sleep behavior, but cannot explain how.	*"I shaded this area because after 1 day in London the person will still be on the same sleep schedule as they would in Denver, CO. This is due to their biological clock."*
3	Alludes to hidden physiological mechanisms	Some scientific vocabulary is used to suggest cellular/internal control of sleep behavior, but no specific mechanism is described.	*"The person would probably be asleep when it is morning here because their brain wasn't used to the time in England. Jetlag!!"*
2	Accounts restricted to the organismal level	Observable changes occur in direct response to the environment, with no intermediate physiological mechanism.	*"Now your body has changed to London time."*
1	Stories at the organismal level based on personal experience / cultural models	Sleep behavior attributed to conscious effort. Ideas about the body refueling during sleep.	*"You wouldn't be tired if you slept on the plane, so you probably wouldn't go to bed until noon."*
0	No response / unintelligible / negligible	-	-

Table 6. Rubric for scoring students' arguments in the posttest interview. Modified from McNeill et al. (2006).

	Level			
	0	**1**	**2**	**3**
Claim: An assertion that answers the original question.	Does not make a claim.	Makes an inaccurate or inappropriate claim.	Makes an appropriate but incomplete claim.	Makes an accurate and complete claim.
Evidence: Scientific data that supports the claim. Data need to be appropriate and sufficient.	Does not provide evidence.	Provides inappropriate evidence.	Provides appropriate but insufficient evidence.	Provides appropriate and sufficient evidence to support the claim.
Reasoning: A justification that links the claim and evidence, using appropriate and sufficient scientific principles.	Does not provide reasoning.	Reasoning does not link evidence to claim. Scientific principles are missing, vague, or inaccurate. May rely on informal / non-scientific principles.	Reasoning links some of the evidence to the claim. Includes some, but insufficient scientific principles.	Reasoning links multiple forms of evidence to claim. Includes appropriate and sufficient scientific principles.

Table 7. Summary of Hierarchical Regression Analysis for Variables Predicting Student Posttest Score (N = 58).

Variable	B	SE B	B
Step 1			
Pretest score	0.732	0.121	0.630***
Step 2			
Pretest score	0.760	0.118	0.654***
Group	4.250	1.988	0.216*
Step 3			
Pretest score	0.736	0.119	0.640***
Group	3.719	2.032	.188
FRL	-3.958	3.292	-0.124
Step 4			
Pretest score	0.685	0.122	0.596***
Group	3.962	2.009	0.200
FRL	-3.514	3.256	-0.110
Race/ethnicity	-4.177	2.620	-0.168
Step 5			
Pretest score	0.693	0.121	0.603***
Group	4.374	2.016	0.221*
FRL	-3.861	3.241	-.121
Race/ethnicity	-4.180	2.599	-.168
Gender	-2.705	2.006	-.137

Note. R^2 = .397 for Step 1; ΔR^2 = .046 for Step 2 (p < .05); ΔR^2 = .016 for Step 3 (ns); ΔR^2 = .026 for Step 4 (ns); ΔR^2 = .018 for Step 5 (ns). *B* = unstandardized regression coefficient; *SE B* = standard error of *B*; β = standardized regression coefficient; FRL = Free or Reduced Price Lunch. **p* <.05 , ****p* <.001

Table 8. Normalized gain scores by group assignment and demographic variables

Normalized Gain Score	**Commonplace**	**Inquiry**
FRL Status		
does not receive FRL	0.31	0.40
receives FRL	0.26	0.27
Race/Ethnicity		
white	0.32	0.41
non-white	0.20	0.31
Gender		
male	0.34	0.38
female	0.22	0.39

Table 9. Summary of means, standard deviations, effect sizes and confidence intervals for each student achievement and classroom measure.

		Commonplace Group (n=28)		Inquiry-Based Group (n=30)		Effect Size	Effect Size Confidence Interval (95%)	
		Mean	SD	Mean	SD		Lower	Upper
Total Test Scores *(out of 74)*	Pretest	31.11	8.60	29.23	8.49	.47	-.05	0.99
	Posttest	43.61	9.09	46.43	10.57			
	Adjusted Posttest*	42.87		47.12				
Reasoning *(fraction of responses at level 5)*	Pretest	.04	.15	.03	.09	.68	.15	1.21
	Posttest	.14	.19	.27	.28			
	Adjusted Posttest**	.14		.27				
Argumentation *(out of 3)*	Claim	1.59	.45	1.83	.54	.58	.05	1.10
	Adjusted Claim*	1.58		1.84				
	Evidence	1.64	.54	1.98	.57	.74	.21	1.27
	Adjusted Evidence*	1.61		2.01				
	Reasoning	1.59	.54	1.86	.58	.59	.07	1.12
	Adjusted Reasoning*	1.57		1.89				
RTOP *(mean item score, out of 4)*		1.52	.87	3.44	.42	2.81	2.03	3.59
CLES *(out of 85)*		54.26	8.95	61.46	7.75	.86	.32	1.40

* = Posttest scores controlled for the variance in students' total test pretest scores.
** = Posttest scores controlled for variance in students' level 5 reasoning pretest scores.

All effect sizes from adjusted posttest scores are calculated using the unadjusted control group posttest standard deviation (Slavin, 1996).